I'm not robot

Probability and statistics interview questions and answers pdf

Mastering probability and statistics is essential for acing data science interview questions. The job market for data scientists has become increasingly competitive due to the growing popularity of the field. While many focus on programming skills, SQL, and advanced concepts like deep learning and computer vision, statistics and probability are equally important for success in Data Analyst or Product Data Scientist roles. This article will provide crucial concepts and tips for tackling probability and statistics interview questions. Statistics is the backbone of many data analysis tasks, regardless of the industry. In an e-commerce setting, clustering methods rely on statistical techniques to group customers with similar characteristics. Conducting A/B tests in a tech company requires knowledge of hypothesis testing, margin of error, confidence intervals, and sample sizes to determine which product performs better. Even complex deep learning architectures like Variational Autoencoders (VAE) and Bayesian Neural Networks (BNN) are rooted in statistical concepts. Probability Concepts You Should Know for Data Science Interview As a Data Analyst or Data Scientist, statistics plays a vital role in our daily work. Understanding probability concepts is crucial to acing data science interviews. #### Probability Fundamentals Probability is the backbone of statistics, and it's essential that we grasp its basics. This section will cover all probability fundamentals, helping you understand how to apply them to real-world problems. #### Notations and Operations When working with probabilities, various notations and operations come into play. Each notation represents specific events in a probability question and has distinct meanings. Familiarizing yourself with common notations and operations is vital for solving probability questions accurately. * **Union**: The union of two events A and B includes all outcomes from both events. * **Intersection**: The intersection of two events A and B contains only the outcomes common to both events. * **Mutually Exclusive**: Two events are mutually exclusive if none of their outcomes overlap. From a statistical perspective, mutually exclusive events have no overlap. This means that the probability of both events occurring simultaneously is zero. Probability also has some fundamental properties: **Range**: Probability always falls within the range between 0 and 1. If an event's probability is 1, it means the event will definitely happen. Conversely, if an event's probability is 0, it won't happen at all. The closer the probability to 1, the more likely the event will occur. **Complements**: The probability of an event not occurring is equal to the complement of the original event. For example, if there's a 0.8 chance of rain tomorrow, then the probability of no rain would be 1 - 0.8 = 0.2. **Union of Two Events**: The probability of two events occurring together (union) can be calculated using the following formula: However, if the two events are mutually exclusive, their intersection is zero, and the probability of union is simply: In a recent data science interview question from Facebook, the concept of probability union was tested. The question asked: What's the probability of pulling a different color or shape card from a shuffled deck of 52 cards? To solve this problem, we first defined two events: A: pulling a card with a different color B: pulling a card with a different shape Next, we calculated the probabilities for each event: Finally, we applied the formula to calculate the probability of union: This concept is commonly tested in data science interviews. Conditional probabilities can be grasped fairly easily. It refers to the likelihood of an occurrence given that another event has taken place. Mathematically, this is denoted as P(A|B), meaning the probability of A happening given that B has occurred. Dependent events are those where the outcome of one influences the probability of the other. An example would be drawing two cards from a deck, hoping to get a 5 heart. Initially, the chance is 1/52, but after missing it on the first draw, the likelihood for the second attempt drops to 1/51. Conversely, events are considered independent if one does not affect the probability of the other. Tossing a coin twice and hoping to get tails in one of them serves as an example. The probability remains constant at ½ regardless of the first toss's outcome. In this problem, we're looking for the probability of event A (rolling die 2 to get a 4) occurring. To determine this, we first need to establish whether event A is dependent on or independent from another event B (the outcome of rolling die 1). Since the probability of getting a 4 with die 2 doesn't change regardless of the outcome of rolling die 1, we can classify our event as an independent event. With this understanding, we can apply the concept of permutations and combinations. While often used interchangeably, these terms have distinct meanings based on whether order matters. In permutations, order is crucial; in combinations, it's not. A key difference lies in how they're calculated: permutations involve multiplying possible outcomes, whereas combinations use factorials to determine the number of ways items can be selected without regard for order. To illustrate this concept, let's consider a tennis competition with 8 athletes competing for 3 medals (Gold, Silver, and Bronze). In the first scenario where each medal is distinct based on the winner's position, we deal with permutations. We begin by selecting one athlete out of 8 for the Gold medal, then another out of the remaining 7 for the Silver medal, and finally one out of the last 6 for the Bronze medal. This gives us a total of 336 possible combinations (calculated as 8 x 7 x 6). The general formula for permutations is n^k, where n represents the total number of items to choose from, and k is the number of selections. In our tennis example, with 8 athletes and 3 medals, we calculate the permutations as 8^3 = 512. A similar problem was posed in a data science interview: Given three people competing for three places (1st, 2nd, and 3rd), how many different combinations are there? This question is essentially asking for the number of ways these three positions can be filled without regard to order. The solution involves arranging the 3 people into 3 spots, which gives us a total of P(3) = 3! (factorial) permutations, equaling 6 possible arrangements. This distinction between permutations and combinations is fundamental in probability theory, especially when calculating probabilities for independent events or scenarios where order does not matter. In this scenario, we need to calculate the number of ways to choose a subset of items from a larger set, where the order does not matter. This can be achieved using the general equation for combinations. For example, let's say we have a group of three winners: athletes A, B, and C, who will receive souvenirs. Since all three winners receive the same souvenirs, the order doesn't matter, making it a combination problem rather than a permutation problem. The general equation for combinations is C(n,k) = n! / (k!(n-k)!), where n is the total number of items and k is the total number of items to be ordered. There are two key properties that can be derived from probability distributions: the expected value and the variance. The expected value represents the average value of a sample, while the variance measures the theoretical limit of the variability of samples as their size approaches infinity. In this section, we'll focus on some of the most commonly asked probability distributions during an interview and their use cases. Discrete probability distributions are used to visualize categorical data, characterized by a discrete set of possible outcomes with a probability mass function (PMF). The PMF has an x-axis representing the possible outcomes and a y-axis representing the probability of each event. Several discrete probability distributions are worth knowing for data science interview preparations, including Bernoulli distribution. Bernoulli Distribution The Bernoulli distribution has only two possible outcomes: 0 (failure) or 1 (success). The probability of success is p, while the probability of failure is (1-p). In a fair coin toss, the probability of getting a tail (failure) is 0.5, and thus the probability of getting a head (success) is also 0.5. However, the probability of two outcomes does not need to be equal, as seen in an unfair coin where the probability of getting a head is 0.7 and the probability of getting a tail is 0.3. The expected value of a random variable in Bernoulli distribution is, while the variance is. Example Question: Facebook asked about Bernoulli distribution in one of their data science interviews. Here was the question: "Three ants are sitting at the three corners of an equilateral triangle. Each ant randomly picks a direction and starts to move. What is the probability that all three ants will move in the same direction?" The binomial distribution builds upon the Bernoulli distribution, which itself is based on a simple probability concept. To understand this, consider three ants walking along the edge of a triangle, where each ant moves either left (0) or right (1). The question here revolves around calculating the probability that none of these ants collide. In this scenario, each ant's movement is independent and can be viewed as a Bernoulli trial. Therefore, the binomial distribution conceptually extends this idea to more trials—i.e., tossing a coin 100 times instead of once. Out of a hundred coin tosses, how often does it land on its tail side? This example follows a binomial distribution pattern. It measures the chances of success in n trials where each trial is independent of the others, meaning the outcome of one coin toss doesn't affect the next. The binomial distribution equation looks like this: n = number of trials k = number of successes Its expected value and variance are also unique, given by specific formulas. This knowledge is crucial for data science interviews, such as the question Verizon Wireless once asked: "What's the probability of getting at least one 5 when throwing a dice seven times?" To solve this, you find the probability of not getting any fives (k=0) and subtract that from 1. The binomial distribution is closely related to another concept called hypergeometric distribution. The difference lies in how each draw affects the probability of the next. For example, drawing a heart from a deck doesn't change your chances for the next card; it does in a hypergeometric scenario. measure successes k out of n trials, but in this case, we're focusing on how many trials it takes to get our first success. Unlike binomial distribution, where trials happen with replacement, hypergeometric distribution has no replacement. Let's imagine a deck of cards: what's the probability of pulling a different color or shape card? We can define these events as:A: getting a card with a different colorB: getting a card with a different shapeAfter drawing one card, the probability of drawing another card with a different color or shape (no replacement) is the same. Geometric distribution is closely related to binomial distribution and measures the number of trials until we get our first success. You might notice that geometric distribution's probability mass function (PMF) looks different from binomial distribution's PMF, but they share a similar concept. Let's use coin tosses as an example: with geometric distribution, we can estimate how many coin tosses it takes to get our first tail. Binomial distribution measures the number of successes in n trials. Geometric distribution asks: "how many failures until we get the first success?" This difference is why geometric distribution's PMF has a certain shape. As we observe more trials, it becomes less likely that we didn't get any success. The concept of negative binomial distribution arises from scenarios where we're interested in the number of trials until a specific number of successes occur. This is different from the geometric distribution, which focuses on the number of failures before achieving the first success. Consider a scenario where you visit a casino with two dice to play a game that rewards you $10 for rolling a 5. The cost of playing each round is $5. There are four favorable outcomes: {4,1}, {2,3}, {3,2}, and {1,4}. With these outcomes, the probability of winning is 1/9. The expected payout in this scenario can be calculated using the formula for negative binomial distribution, which takes into account the number of trials needed to achieve n successes. In this case, we're looking for a single success (rolling a 5). The expected value is given by: E(X) = np / p where n is the number of successes (1 in this case), p is the probability of success (1/9), and E(X) is the expected value. Expected outcomes can be modelled using continuous probability distributions, which are different from discrete ones. These distributions use a probability density function (PDF) instead of probability mass functions (PMF). PDF is the area under a curve between two points, as shown in the figure below. Normal distributions have a bell-curved shape and are widely used due to their properties. The mean and standard deviation are important parameters for this distribution. The empirical rule defines normal distribution, stating that roughly 68% of observations fall within one standard deviation from the mean, 95% within two, and 99.7% within three. The standard normal distribution has a mean value of zero and variance of one. This is also known as a z-distribution, used for comparing sample averages to population averages when the population standard deviation is known. The student t-distribution has a similar shape but with heavier tails than the normal distribution, depending on the degree of freedom (DOF). When exploring data distributions, understanding normalcy is crucial. This type of distribution is often used when testing whether a sample average corresponds to the population average when the population standard deviation is unknown. The student t-test will be covered in-depth later. To kick-start our statistical journey, let's dive into basic statistics concepts. We'll focus on measures of center: mean, median, and mode. These metrics quantify the central tendency of data, indicating where the data's center lies. The key difference between them lies in their calculation methods. **Mean:** The average value of all data points. **Median:** The middle value when data is sorted in order. **Mode:** The most frequently occurring number. To visualize these differences, let's examine datasets with varying distributions. When data is normally distributed, mean, median, and mode coincide. However, if the distribution is left-skewed or right-skewed, the mean can be affected by outliers. The mean is a sensitive metric, making it essential to check for outliers before using it as a center-measuring tool. A notable example of this concept was asked in a Facebook data science interview: "In Mexico, which age would be higher, the mean or median, and why?" To answer this, we need to determine the age distribution shape. Statista reveals that Mexico's age distribution is right-skewed between 2010-2020. Therefore, we can conclude that the mean age is higher than the median. This is just a taste of basic statistics concepts crucial for data science interviews. Let's continue our journey by exploring other essential statistical ideas! The question "How would you impute missing information?" asked by Airbnb in a data science interview seems straightforward but requires an understanding of statistical concepts such as means and medians for numerical data. When dealing with missing values, it's crucial to check the distribution of the data first. If there are outliers, using the mean could introduce bias, making the median a better choice since it's more robust against outliers. For categorical data, imputing with the mode is sensible as it fills missing values with the most frequent categories. Moving on to measures of spread, these metrics show how spread out our data is. There are four main types: range, variance, standard deviation, and quartiles. The simplest measure is the range, which is calculated by subtracting the lowest from the highest value in a dataset, giving a spread of $70,000 if the lowest salary is $35,000 and the highest is $105,000. However, variance measures dispersion around the mean but is less intuitive due to being in squared units. The formula for variance is S = Σ(xi - x_bar)^2 / n, where S is the variance, xi is each observation value, x_bar is the mean of all observations, and n is the total number of observations. Standard deviation solves this unit problem by also measuring dispersion around the mean but in the same units as the data. This makes it more intuitive than variance for defining how spread out data is from its average. The standard deviation, which gauges the variability in a dataset, can be seen as a fine-tuning of variance. It serves as a unit of measurement for the spread of data points from the mean value, making it easier to quantify this dispersion. For normally distributed data – that is, data resembling a perfect bell-shaped curve – standard deviation plays a significant role in approximating the proportion of data points according to their proximity to the mean. An example often cited involves the distribution of employee salaries around a mean annual salary of $60,000. In such scenarios, approximately 68% of salaries lie within one standard deviation from the mean, while about 95% and 99.7% of salaries fall within two and three standard deviations, respectively. Another tool for measuring spread is quartiles, which divide data into four parts: the lower quartile (25%), the interquartile range (25-50% and 50-75%), and the upper quartile (75%). Visualizing these concepts through boxplots can simplify understanding. In a data science interview, questions about measures of spread are common, such as identifying outliers. Detecting an outlier involves identifying whether a new observation lies within the norm of the distribution. If data is normally distributed, lying beyond three standard deviations from the mean often indicates an outlier. Boxplots are also useful for visualizing this – any data point outside the maximum and minimum values of a boxplot can be considered anomalous. Hypothesis Testing in Inferential Statistics Inferential statistics involves testing assumptions or beliefs about a general population using sample data. Hypothesis testing is a crucial concept in this field, allowing researchers to determine which hypothesis is more plausible. **Null and Alternative Hypotheses** A null hypothesis assumes that everything is fine with the current situation, while an alternative hypothesis proposes a counter narrative. For example, if you want to know if there's a significant difference between the IQ of men and female students in a school, the hypotheses would be: * Null hypothesis: The IQ of men and female students are the same * Alternative hypothesis: The IQ of mean and female students are different **Significance Level** The significance level is a critical concept that determines our tolerance for rejecting a null hypothesis. It can vary depending on one's interest, but a common rule of thumb is a 5% or 0.05 significance level. A lower significance level means we're more cautious in rejecting the null hypothesis. **Confidence Interval** The confidence interval represents the plausible range of values for a population parameter. In two-tailed hypothesis testing, the confidence interval complements the significance level, meaning that if our significance level is 5%, then the confidence interval would be 95%. Computing confidence intervals involves considering sample size, standard deviation, and confidence levels to estimate a population parameter. The equation for calculating confidence intervals is based on sample mean (x_bar), Z-score corresponding to the desired confidence level, sample standard deviation (σ), and sample size (n). A 95% confidence interval corresponds to a Z-score of 1.96, reflecting the degree of uncertainty in the estimate. Your observation suggests that if the null hypothesis is considered true, combining a p-value with a significance level provides insight into whether to reject it in hypothesis testing. For instance, if the p-value from your observation is 0.02 and the significance level is 0.05, you would reject the null hypothesis since the p-value is less than the significance level. Conversely, if the p-value is 0.02 but the significance level is 0.01, you cannot reject the null hypothesis. One-Sample Z-Test ========== A One-Sample Z-Test is used to compare the mean of a sample data with the known population mean. The test assumes that the population standard deviation is known. Z Value Calculation ---------- * `x_bar`: Sample mean * `μ`: Population mean * `σ`: Standard deviation of population * `n`: Number of observations The Z value can be used to look up the corresponding p-value using a z-table or statistical libraries in Python or R. One-Sample t-Test for Population Mean ---------- This test is used when we want to check if the mean of our sample data corresponds to the population mean. The assumption here is that we don't know the population standard deviation, which is more realistic in real-life scenarios. * Sample must be randomly selected * If sample size is very small (around 15), then the data should have a normal distribution * If sample size is considerably large (around 40), then the test is safe to conduct even when the data is skewed One-Sample t-Test Formula ---------- * `x_bar`: Sample mean * `μ`: Population mean * `s`: Standard deviation of samples * `n`: Number of observations The intuition behind this test is similar to the One-Sample Z-Test, but it uses the sample standard deviation instead of knowing the population standard deviation. The degree of freedom (DOF) for this test is `n - 1`. Paired t-Test ---------- This test is used when we want to check if a treatment yields a statistically different result within the same sample. The conditions to fulfill are generally the same as the one-sample t-test. The formula for this test is also the same, but the interpretation of the result differs. Instead of measuring the mean of our variable of interest, we're more interested in the mean or standard deviation difference of that variable. Two-Sample t-Test ---------- This test is used when we want to check if two different samples with different treatments yield a statistically different result between those samples. The conditions to fulfill are the same as the one-sample t-test, and it's assumed that the distribution of both samples is similar. The result will be more reliable if the size of the sample is larger. When dealing with samples, understanding the t-statistic's intuition is crucial. Essentially, it measures the difference in mean and standard deviation between two or more samples. This concept is similar to that of a paired t-test. Now, let's move on to ANOVA (Analysis of Variance). When you have multiple treatments applied to different samples and want to determine if they yield statistically significant results, ANOVA comes into play. The conditions required are the same as those for the two-sample t-test, but with more than two samples. To perform an ANOVA test: 1. Calculate the overall mean of all samples. 2. Determine the within-group deviation by subtracting each sample's mean from its respective members' values. 3. Calculate the between-group variation by comparing each sample to the total mean of all samples. 4. Use these results to compute the F-statistic, which is the ratio of the between-group variation to the within-group deviation. Keep in mind that ANOVA can be time-consuming if done manually; therefore, it's usually performed with statistical software or libraries in a programming language. For instance, you could use Python to solve Probability and Statistics questions. Check out our post on 'Probability and Statistics Interview Questions and Answers' for more information. Another important test is the Chi-Square Goodness of Fit. This test helps determine if sample data accurately represents the full population. Instead of focusing on means, we look at proportions. When to use this test: If you want to test whether a sample's proportion matches that of the entire population. For instance, let's say you have 10 packs of M&M chocolate and you're trying to determine if the proportion of different colors in each pack is equal. Conditions for the Chi-Square Goodness of Fit: 1. The sample should be randomly selected. 2. Data should be categorical or nominal. 3. The sample size should be large enough to have at least 5 items per category. The formula for the Chi-Square Goodness of Fit test is: where O is the observed frequency count, E is the expected frequency count, and k is the total number of categories. After calculating the Goodness of Fit, you'll receive a p-Value, similar to other tests mentioned above. The Chi-Square Test of Independence provides insight into whether two or more sample data are related or not. This test is similar to ANOVA, but instead of focusing on means, it examines proportions. For instance, if you own a cinema and want to know if the movie genre affects snack sales, this test can help. The conditions for this test include having two or more samples, just like in the Goodness of Fit test. The formula for the Chi-Square Test of Independence involves calculating the p-Value, which helps decide whether to reject the null hypothesis. In an example scenario, Amazon asked how to check if assignment to different buckets was truly random in an A/B test. Since this is a categorical variable test, the Chi-Square Test of Independence can be used to compare variables A and B. To conduct hypothesis testing, follow these steps: First, establish the null and alternative hypotheses. Next, choose your significance level. Then, select a suitable statistical test based on your observations. Calculate the p-Value from this test. Finally, make a decision based on the p-Value. We must decide whether to reject the null hypothesis or not. Facebook asked about a concept called inferential statistics in an interview. The question was: "If a product is being used differently by two groups, what do you think is happening and how would you test it?" To analyze this situation, we gather data first. Then we use the two-sample t-test to see if there's a difference between the two groups. Our hypothesis has two parts: - The null hypothesis says that the average rate of use for both groups is the same. - The alternative hypothesis says that the average rate of use for both groups is different. We choose a significance level, usually 0.05. We calculate the average and standard deviation for each group using the data. Then we calculate the t-statistic using the two-sample t-test. If the t-statistic is less than our significance level (0.05), we reject the null hypothesis and say that the average rate of use for both groups is different. If the t-statistic is more than our significance level, we stick with the null hypothesis and say that the average rate of use for both groups is similar. There's always uncertainty in statistical testing. The main source of this uncertainty is our chosen significance level and the quality of the sample data. This means we can make two types of errors when conducting hypothesis testing: - Type I error: Falsely rejecting the null hypothesis. - Type II error: Failing to reject the null hypothesis. Statistical power is the probability of correctly rejecting the null hypothesis. It's also the probability of making fewer Type II errors. The higher the statistical power, the lower the risk of a Type II error. Data quality issues often mean you lack sufficient sample data to produce a convincing p-value for rejecting or failing to reject the null hypothesis. This raises the question: what is the minimum sample size required to ensure a convincing p-value? In power analysis, we consider four key parameters: effect size, sample size, significance level, and statistical power. You already know the latter three, but the former requires calculating using statistical measures like Pearson's correlation or Cohen's d. To conduct a power analysis, you need to predefine the values of effect size, significance level, and statistical power. Default values include 0.05 for significance level, 0.8 for large effect sizes (Cohen's d), and 0.8 for statistical power. However, these can be adjusted depending on your research area. Due to the complexity of manual computation, power analysis is often implemented in programming languages using statistical libraries. Now, let's move on to Bayes' theorem, which offers an alternative perspective on inferential statistics. Bayes' theorem applies probability theory to update a prior belief about an event as new evidence emerges. It produces a posterior belief. The fundamental concept behind this theorem is known as Bayes' rule, similar to conditional probability. Here's the equation for Bayes' rule: P(A|B) = (P(B|A) * P(A)) / P(B) In this formula, A and B represent events. We can define P(A|B) as the probability of event A given that event B has occurred. Similarly, we can define P(B|A) as the probability of event B given that event A has occurred. The probability of event B occurring given that event A has occurred is a fundamental concept in statistics. The Bayes' rule illustrates this idea by highlighting the importance of prior, likelihood, marginal, and posterior probabilities. To further explain these concepts, let's examine the questions they attempt to answer: * Posterior: What is the probability of our hypothesis being true given the observed evidence? * Likelihood: How probable is the evidence if our hypothesis is true? * Prior: What is the probability of our hypothesis before observing any evidence? * Marginal: What is the probability of the evidence under all possible hypotheses? For instance, Facebook asked a data science interview question about Bayes' theorem: "You're about to board a flight to Seattle and want to know if you should bring an umbrella. You call three random friends living there and ask each independently if it's raining. Each friend has a 2/3 chance of telling the truth and a 1/3 chance of lying. All three friends tell you that 'Yes,' it is raining. What is the probability that it's actually raining in Seattle?" To solve this problem, let's assume various events: * A: It's raining in Seattle * A': It's not raining in Seattle * Xi: A random variable with a Bernoulli distribution, where Xi = 1 if a friend says "Yes," and Xi = 0 if a friend says "No." Since we don't know the probability of rain in Seattle, let's assume it to be 0.5. The problem states that all three friends said "Yes," it will be raining in Seattle. We can express this as: We know that the probability of our friends telling the truth is 2/3. We can express the probability of a friend saying "Yes" given that it's raining in Seattle as: and the probability of a friend saying "No" given that it's not raining in Seattle as: Now that we have the necessary information, we can find the probability of rain in Seattle given that our three friends say "Yes" using Bayes' theorem. When preparing for probability and statistics interview questions, it's essential to understand these concepts thoroughly. To conquer your probability and statistics interview questions, keep these tips in mind. Firstly, focus on learning the concepts rather than just memorizing formulas. This will enable you to approach each question correctly, regardless of whether it involves binomial, geometric, or negative binomial distributions. For instance, when asked about the probability of 5 tails in 20 coin tosses, knowing that it can be solved with a binomial distribution is key. Secondly, use pen and paper to brainstorm your thoughts when dealing with math-heavy questions like these. This will help you solve problems step-by-step and think more concisely about the problem at hand. Lastly, don't try to learn every concept in one sitting - mastering statistics takes consistent effort over time. Set aside an hour each day to practice solving probability and statistics interview questions, and make sure to take breaks as well. With time and consistency, you'll become proficient in recognizing recurring patterns and applying the right concepts to solve each question effectively. When tackling probability and statistics interview questions, it's essential to ensure a deep understanding of the problem being presented. To guarantee comprehension, consider paraphrasing the question in your own words and then verifying with the interviewer if your interpretation aligns with their query. This technique serves two purposes: validating your understanding of the question and affording you a brief moment to contemplate the solution.